DATA ANALYSIS AND VISUALIZATION

Prepared by: 23BCE194(SARTH NAROLA) / 23BCE195(NEVIL ANGHAN)

Course / Course code: Data Analysis and Visualisation / 3CS103ME24

PRACTICAL : 1

**Domain Overview – Education**

The education domain deals with data related to students, academic performance, institutional processes, and learning behavior. Analyzing such data helps institutions improve teaching quality, reduce dropout rates, and personalize education.

**Selected Datasets**

We selected 5 datasets under the education domain to comprehensively analyze various aspects of academic success, dropout trends, performance, technology adoption, and global migration for studies.

| **S.No** | **Dataset Name** | **Rows** | **Features** | **Brief Description** |
| --- | --- | --- | --- | --- |
| 1 | **Education Career Success** | 5000 | 20 | Tracks career outcome based on academic, internship, skill, and GPA-based features |
| 2 | **Student Performance** | 5000 | 33 | Performance indicators including grades, attendance, habits, and family info |
| 3 | **Dropout & Academic Success** | 4424 | 38 | Academic and demographic data to predict student status (Graduate, Dropout, Enrolled) |
| 4 | **Virtual Reality in Education** | 5000 | 20 | Explores the impact of VR usage on engagement and learning outcomes |
| 5 | **Global Student Migration** | 5000 | 20 | Analyzes student movement across countries and their academic/career success |

***Educational Data Analysis for Student Status Prediction***

**Dataset Exploration**

Total Records: 4,424 student entries

Total Features: 37 input features + 1 target variable

Target Variable: Target – Student academic status:

* Dropout
* Graduate
* Enrolled

**Applications of the Dataset**

1. **Machine Learning Tasks**

* **Classification**:
  + Predict final status (Graduate, Dropout, Enrolled)
  + Identify risk groups based on features
* **Regression**:
  + Predict continuous scores (e.g., grades, average performance)

1. **Practical Use-Cases**

* **Dropout Risk Prediction** – Early detection of high-risk students
* **Scholarship Eligibility** – Based on financial and academic data
* **Performance Monitoring** – Using grade-related features
* **Admission Planning** – Via insights from Application mode and Order
* **Curriculum Optimization** – Through evaluation/approval rates

**Challenges in the Dataset**

| Challenge | Description |
| --- | --- |
| Missing Data | Some features may have null values that must be imputed |
| Imbalanced Target Classes | ‘Graduate’ has more examples than ‘Dropout’ |
| Interpretability | Some features (like Application mode) need encoding |
| Outliers | Extreme values in grades, evaluations may skew results |
| Static Data | Dataset lacks time-series dimension for behavior tracking |

**Stakeholders & Their Use Cases**

* **Institutions** use the data to identify students at risk, optimize scholarship allocation, and better manage academic resources.
* **Teachers** benefit by adjusting their teaching strategies based on student performance analytics.
* **Parents** can monitor their children’s academic progress and receive alerts if any risks are detected.
* **Policy Makers** rely on trends revealed through data to allocate funding effectively and design responsive education policies.

**Key Takeaways**

* The dataset enables critical analysis of **student academic lifecycle**.
* Proper preprocessing (handling missing values, encoding) is essential.
* Predictive modeling can aid **early interventions** for dropout prevention.
* Classification and regression tasks both have strong applicability